Wie sehr können maschinelle Indexierung und modernes Information Retrieval Bibliotheksrecherchen verbessern?

Dipl.-Inf.wiss. Manfred Hauer M.A.
FHS Burgenland, Informationsberufe &
AGI – Information Management Consultants
Neustadt an der Weinstraße
http://www.agi-imc.de
http://www.dandelon.com
Manfred.Hauer@agi-imc.de

Zusammenfassung
Web-OPACs mit Human-Indexierung fallen in einem Retrieval-Test deutlich hinter maschinelle Verfahren zurück. intelligentCAPTURE saugt Content über Scanning, File-Import und Web-Spidering ein und indexiert nach morphosyntaktischen Verfahren. Neben Bibliothekssystemen übernimmt dandelon.com den Content und die Indexate. Dandelon.com ist öffentlich und kostenlos zugänglich für Endbenutzer, Austauschzentrale und Katalogerweiterung für angeschlossene Bibliotheken. Die Kosten sind gegenüber der Humanerschließung wesentlich geringer bei zugleich deutlich höherem Wirkungsgrad in der Recherche.

Abstract
In a benchmarking between human indexing in library catalogs and machine indexing in the open service dandelon.com the machine approach succeeded. It is based on intelligentCAPTURE, an program capturing content via scanning, file import or web spidering. Text is indexed by a build in morphosyntactical engine. Results are pasted to library catalogs as well as to dandelon.com. It is a free of charge service for everybody and exchange center for linked libraries. Cost are much below the human indexing approach but with much better retrieval results.

Reicht ein Web-OPAC im Zeitalter des Internets noch aus? Ca. 90 % aller Bibliotheksrecherchen durch Benutzer sind Themenrecherchen. Ein Anteil dieser Recherchen bringt kein Ergebnis. Es kann leicht gemessen werden, dass null Medien gefunden wurden. Die Gründe hierfür wurden auch immer wieder untersucht: Plural- anstelle Singularformen, zu spezifische Suchbegriffe, Schreib- oder Bedienungsfehler. Zu wenig untersucht sind aber die Recherchen, die nicht mit einer Ausleihe enden, denn auch dann kann man in vielen Fällen von einem Retrieval-Mangel ausgehen. Schließlich: Von den ausgeliehenen Büchern werden nach Einschätzung vieler Bibliothekare 80 % nicht weiter als bis zum Inhaltsverzeichnis gelesen (außer in Präsenzbibliotheken) - und erst nach Wochen zurückgegeben. Ein Politiker würde dies neudeutsch als „ein Vermittlungsproblem" bezeichnen. Ein Controller als nicht hinreichende Kapitalnutzung. Einfacher machen es sich immer mehr Studenten und Wissenschaftler, ihr Wissensaustausch vollzieht sich zunehmend an anderen Orten.

Bibliotheken (als Funktion) sind unverzichtbar für die wissenschaftliche Kommunikation. Deshalb geht es darum, Wege zu finden und auch zu beschreiten, welche die Schätze von Bibliotheken (als Institution) effizienter an die Zielgruppe bringen. Der Einsatz von Information Retrieval-Technologie, neue Erschließungsmethoden und neuer Content sind Ansätze dazu.

Bibliometrischer Retrieval-Test: OPAC gegen dandelon.com

Im Rahmen einer Lehrveranstaltung „Information Retrieval in Bibliotheken" mit Studenten der FH Darmstadt wurden empirisch die Wirkungen des Retrieval-Systems FAST an der UB Bielefeld und von der Content-Generierung und Indexierungsmaschine intelligentCAPTURE an der Vorarlberger Landesbibliothek und in den Daten der Vorarlberger Landesbibliothek unter dandelon.com in einem ersten Experiment evaluiert. Weitere klassische Universitätskataloge, Verbundkataloge und KVK wurden im Vergleich dazu genutzt. Über die Eingabe von beliebigen Vornamen wurden Titel in dandelon.com herausgesucht und diese Listen solange mit den anderen Katalogen verglichen, bis ein

Titel gefunden wurde, der in allen Katalogen vorkam. Dazu braucht man ca. 10 bis 15 Anläufe, der Grad der Überdeckung von Bibliotheken ist also eher gering (das spricht für übergreifende, statt lokaler Kataloge und räumlich unabhängige Ausleihsysteme). Schließlich wurden drei Bücher gefunden. Je ein Student studierte intensiv das Inhaltsverzeichnis - glücklicherweise waren alle drei Bücher hinreichend allgemein verständlich - und er gab den Mitstudenten den Auftrag, genau dieses Buch in allen Katalogen zu suchen. Jedes Buch wurde dazu ausführlich referiert. Je Katalog waren je 10 Versuche erlaubt und der gesuchte Titel sollte auf den ersten beiden Bildschirmseiten auftauchen. Diese Situation ist mit der eines Wissenschaftlers, der sein Thema genau kennt, doch nicht genau das Ziel, gut zu vergleichen.
Normale OPACs auch von den Verbundkatalogen fielen bei den Recherchen fast immer durch, leicht besser schneidet Bielefeld ab, weil durch die technische Aufbereitung der Katalogdaten durch das Retrieval-System FAST Ähnlichkeit und Ranking möglich sind. Doch auch dort erwies sich die klassische, aber doch sehr kurze Indexierung als problematisch. Eine etwas breitere Human-Indexierung an der Vorarlberger Landesbibliothek wirkte sich deutlich positiver aus, noch wirksamer sind die zusätzlichen maschinell generierten Indexate, die in ALEPH ergänzt wurden. Doch der reiche Content, die Indexierung und die automatische Thesaurusunterstützung in dandelon.com brachte die gesuchten drei Titel fast bei jedem Studenten beim ersten oder zweiten Versuch über Ranking sortiert auf die erste Bildschirmseite. Ein sehr klares Ergebnis zugunsten maschineller Indexierung und semantischen Retrievals.

Dieser kleine Test sollte empirisch deutlich weiter ausgearbeitet werden, doch er zeigt schon im Ansatz, dass auch bei sehr genauer Kenntnis eines Thema eine systematische treffsichere Suche in OPACs bis heute nicht möglich ist. Die Indexierungssprachen sind zu grob, der Indexierungsumfang zu eng, die Menge der indexierten Titel zu niedrig, Volltexte nicht suchbar. Obwohl alle Studenten vier oder mehr Semester Informations- und Bibliothekswissenschaften hinter sich hatten und durchaus geschickt und professionell recherchierten, konnte man im Ansatz nicht mehr aus diesen klassischen Katalogen herausholen.

(Die Bibliothekare in Industrieunternehmen arbeiten meist ähnlich wie in öffentliche Bibliotheken - und werden derzeit reihenweise von Forschungsleitern und Controllern abgeschafft, obwohl dort die Wertschöpfung aus Wissen stets viel höher ist als in der öffentlichen Wissenschaft. Rote Lampe!)

Diese Erkenntnis ist nun für Insider nicht wirklich neu. Nur bislang konnte man sich hinter dem allgemeinen „State-of-the-Art“ gut verstecken. Doch warum ist dandelon.com wesentlich effizienter als alle getesteten OPACs? Die Antwort ist einfach: die maschinelle Indexierung in intelligentCAPTURE generiert ein Indexat, das wesentlich mehr suchbare Worte aus einem Text herausholt, als Sacherschließer (sofern sie überhaupt noch indexieren - „sollen es doch andere Verbundteilnehmer tun ..."), die Worte werden automatisch auf die Grundform reduziert und über Thesauri Suchworte ergänzt.

Vom Projekt zum Produkt intelligentCAPTURE

Wegen dieses Mangels hat die Vorarlberger Landesbibliothek zusammen mit AGI - Information Management Consultants ein Projekt vor über zwei Jahren begonnen, das über den Begriff „Kataloganreicherung" deutlich hinausgeht. Daraus entstand das Produkt intelligentCAPTURE, dann vor einem Jahr das Produkt intelligentSEARCH und seit Frühjahr 2004 der öffentliche Service dandelon.com.

Hinter dem internationalen wissenschaftlichen Portal steckt intelligentSEARCH und das Produktionssystem intelligentCAPTURE. intelligentCAPTURE versteht sich als „Durchlauferhitzer", um Content „heiß" zu machen und an beliebige Zielsysteme zu übergeben. Fast 20.000 Bücher sind in dandelon.com derzeit mittels Scanning, OCR und PDF-Konvertierung aufbereitet und der Text des jeweiligen Dokumentes maschinell mit der integrierten CAI-Engine (Computer Aided Indexing) inhaltlich indexiert worden. Durch morphosyntaktische, semantische, heuristische und statistische Verfahren der Textanalyse werden inhaltsbeschreibende Metadaten zu den jeweiligen Dokumenten ergänzt. intelligentCAPTURE übergibt diese Metadaten an Bibliotheksmanagementsysteme, die meist auf relationalen Datenbank-Management-Systemen beruhen. Dadurch kann auch in diesen OPACs, den elektronischen Bibliothekskatalogen, besser recherchiert werden und es können die Inhaltsverzeichnisse von Büchern, Texte von Aufsätzen angezeigt werden. Mittlerweile hat intelligentCAPTURE nicht nur einen Scan/OCR-Workflow, sondern erkennt auch automatisch ob Image-Dateien, PDF-Images oder eigentliche PDFs als Dateien angeliefert werden. Jeweils wird der

Text richtig extrahiert - auch bei mehrspaltigen Texten. Bei der Textextraktion können einzelne Textbereiche wie Zusammenfassungen oder Dokumentenschlüssel wie die internationale DOI erkannt und extrahiert werden. Wer aber seine zu verarbeitenden Texte noch gar nicht hat, aber weiß, wo diese im Web oder Intranet zu holen sind, kann eine URL manuell eingeben und entsprechende Spider-Settings wählen und periodisch wiederholen lassen. Das geht auch mit Listen: So wird z.B. die Zeitschriftenliste mit Links auf Artikel von eJournals von Swets Blackwell wöchentlich hereingeholt und die Artikel im Internet gespidert, indexiert und das Indexat gespeichert. Spidering von Open Archives stehen auf der Agenda. Ähnliches ist mit Forschungsinstituten in Entwicklung. Die große Stärke von intelligentCAPTURE besteht in dem hohen Maß an Automation durch parallel laufende Workflow-Unterprogramme, die auch über mehrere Rechner verteilt ablaufen können. 2005 wird wohl noch Spracherkennung ergänzt, um textlose Objekte effizient zu indexieren: Bilder, Pläne, Gegenstände, Videos. Das Trägersystem IBM Lotus Notes & Domino war schon immer multimedial – bis hin zu Videostreaming ist alles machbar.

intelligentCAPTURE hat eine CAI-Version für allgemeine Bibliotheken und kann zusätzlich für spezielle Domänen/Themenfelder aufbereitet werden. Für Medizin, Technik, Wirtschaft und andere stehen domänenspezifische CAI-Versionen bereit. Bei der Texterkennung fallen laufend neue Begriffe auf, die über die Dokumentenkollektion hinweg statistisch ausgewertet werden können und dann intellektuell Eingang in die Klassifikationen, Thesauri, Topic Maps, semantischen Netze finden. IC INDEX von AGI hat sich hier seit über 20 Jahren einschlägig bewährt für Thesaurus und Klassifikation. Mehrsprachigkeit der Netze ist ein weiteres Highlight, hier steckt maschinelle Übersetzungstechnologie mit dahinter. IC INDEX lässt sich nicht nur mit der Human-Indexierung und verschiedenen Retrieval-Systemen koppeln, sondern kann seine Netze auch direkt an die CAI-Engine exportieren. Die Stärke der maschinellen Indexierung kommt jeweils aus der implementierten Linguistik und Semantik. Analog ist es bei Menschen: gut ausgebildete, erfahrene, sprachlich gewandte Menschen sind in ihrer Domäne „Unwissenden“ überlegen.

Dandelon.com als Portal und Verteilzentrum

intelligentSEARCH kam als Idee erst auf, als sich zeigte, dass die Bibliothekssysteme im Vergleich zu intelligentCAPTURE stets schlechtere Suchergebnisse bei inhaltlicher Suche zeigten. Auch in intelligentCAPTURE ist eine Standard-Suchfunktion verfügbar. intelligentCAPTURE und auch die anderen Produkte basiert auf IBM Lotus Notes & Domino, wo die GTR (Global Text Retrieval) immer enthalten ist. Sie ist eine n-Gram-Retrieval-Engine mit Stemming, Fuzzy-Search, Feldsuche, numerische Suche, Datumssuche, Termweighting und kann über mehrere Domino-Datenbanken optional gleichzeitig suchen. intelligentSEARCH nutzt die GTR-Funktionen geht aber darüber hinaus. So sind bislang über 360.000 Fachbegriffe aus verschiedenen Themenfeldern in Form von semantischen Netzstrukturen zur Suche parallel geschaltet - in IC INDEX-Datenbanken. Daraus resultiert eine automatische Erweiterung und teilweise auch Übersetzung der Anfrage. Optional kann jeder Benutzer diese „Topic Maps" über Flash-Visualisierung auch ansehen, darin navigieren und für die Suche gezielt aussuchen. Da meist im Deutschen die Grundformen - also Singular-Formen - in der Indexierung abgelegt werden, konvertiert ein Programm im Retrieval vorab Plurale oder andere Wortformen in die Grundform. Auch das Stemming von GTR hat das gleiche Ziel, doch nicht den gleichen Wirkungsgrad im Deutschen. Für Englisch reicht das Stemming. Für die Ergebnisanzeige wurde ein Highlighting aller Suchbegriffe ergänzt und bei der Dokumentenanzeige öffnen sich automatisch Metadaten und die Dokumente.
Weil nicht jedes Buch einer Bibliothek auch ad hoc ausleihbar ist oder die eigene Bibliothek das Buch nicht hat, gibt es eine eBusiness-Lösung. Ein Hintergrundagent ermittelt den aktuellen Preis und sendet eine Bestellung an „Missing Link“ in Bremen. Diese Versandbuchhandlung liefert jedes international verfügbare Buch oder Medium überall hin. Vergriffene Titel bzw. Seiten daraus können die Bibliotheken über eine integrierte Document-Delivery-Funktion (Scan/Send) in intelligentCAPTURE leicht übermitteln.

Dandelon.com als Service für Endbenutzer basiert auf intelligentSEARCH und ist darüber hinaus zugleich ein internationales Verbundzentrum, Austauschzentrum für Inhalte, die andere Bibliotheken schon erschlossen haben. intelligentCAPTURE prüft automatisch, welche Inhalte schon im zentralen Pool verfügbar sind und beliefert damit die anderen Bibliotheken, sobald dort der gleiche Titel eintrifft. Jede angeschlossene Bibliothek wird ab Herbst dandelon.com als zusätzliches Front-End zum eigenen Katalog nutzen können – kommt der Suchende aus einem OPAC, sucht dandelon.com auch nur im maschinell erschlossenen Bestand dieser Bibliothek. Für die Vorarlberger Landesbibliothek ist dies schon implementiert.

Alternativ kann intelligentSEARCH als Software natürlich auch für andere Portale genutzt werden - das „Portal Informationswissenschaft“ ist ein erstes Beispiel (www.dgi-info.de).

90 % weniger Kosten bei der Erschließung

Die Vorarlberger Landesbibliothek hat über 2 Jahre gerechnet inzwischen Kosten von über 2 EURO pro gescanntem Buch (Personal, Hard- und Software). Mit dem Spidering von eJournals und Forschungsinstituten, die demnächst weitgehend automatisiert ablaufen werden, fallen die Stückkosten deutlich. Human-Erschließung kostet ca. 16 EURO, ist also sehr viel teurer und wie die Retrieval-Tests zeigten, wesentlich weniger wirksam. Den neuen Content (eJournals, Open Archives, Websites, Blogs, Forschungsinstitute) erreicht die Human-Indexierung in Bibliotheken ohnehin nicht mehr. Doch das eine muss das andere nicht ausschließen – in der Vorarlberger Landesbibliothek ergänzt das eine das andere: mit gutem Erfolg.